WILL MORTALITY RATE OF HIV-INFECTED PATIENTS DECREASE AFTER STARTING ANTIRETROVIRAL THERAPY (ART)?

Shaher Bahakeem

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Richard M. Fairbanks School of Biostatistics,
Indiana University

July 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

_____
Constantin Yiannoutsos, PhD, Chair

_____
Giorgos Bakoyannis, PhD

_____
William Fadel, PhD

Shaher Bahakeem

WILL MORTALITY RATE OF HIV-INFECTED PATIENTS DECREASE AFTER STARTING ANTIRETROVIRAL THERAPY (ART)?

Background: Many authors have indicated that HIV-infected patients mortality risk is higher immediately following the start of Antiretroviral Therapy. However, mortality rate of HIV-infected patients is expected to decrease after starting Antiretroviral Therapy (ART) potentially complicating accurate statistical estimation of patient survival and, more generally, effective monitoring of the evolution of the worldwide epidemic.

Method: In this thesis, we determine if mortality of HIV-patients increases or decreases after the initiation of ART therapy using flexible survival modelling techniques. To achieve this objective, this study uses semi-parametric statistical models for fitting and estimating survival time using different covariates. A combination of the Weibull distribution with splines is compared to the usual Weibull, exponential, and gamma distribution parametric models, and the Cox semi-parametric model. The objective of this study is to compare these models to find the best fitting model so that it can then be used to improve modeling of the survival time and explore the pattern of change in mortality rates for a cohort of HIV-infected patients recruited in a care and treatment program in Uganda.

Results: The analysis shows that flexible survival Weibull models are better than usualoff-parametric and semi-parametric model fitting according to the AIC criterion.

Conclusion: The mortality of HIV-patients is high right after the initiation of ART therapy and decreases rapidly subsequently.

Constantin Yiannoutsos, PhD, Chair

TABLE OF CONTENTS

List of Tables vi
List of Figures vii
Chapter One: Background 1
Chapter Two: Methods 3
    Some Existing Models 3
        The Weibull Model 3
        The Generalized Gamma Model 4
        The Cox Model 5
    Proposed Models 5
        The Parametric Flexible Survival Models 5
        The Royston and Parmar Spline Model 7
Chapter Three: Results 8
Chapter Four: Conclusions 13
References 14
Curriculum Vitae

LIST OF TABLES

Table 1: Comparison of conventional models with Flexible survival models (without covariates) 8
Table 2: Comparison of off shelf models with Flexible survival models (with covariates).....10
Table 3: Summary of the best model (proportional odds models with one internal spline knot).....12

LIST OF FIGURES

Figure 1: Kaplan Meier (step function) patient survival (smooth curve) versus conventional models with Flexible survival models (without covariates).....9
Figure 2: Kaplan Meier (step function) patient survival (smooth curve) versus conventional models with Flexible survival models (with covariates).....11

Chapter One: Background

Several authors like Stringer and colleagues (Stringer, et al., 2006) and May and colleagues (May, et al., 2016) among others have indicated that HIV-infected patients' mortality risk is higher immediately following the start of Antiretroviral Therapy (ART). However, the mortality rate of HIV-infected patients is expected to decrease after starting Antiretroviral Therapy (ART). This fact poses challenge in accurate statistical estimation of patient survival and complicates effective monitoring of the evolution of the worldwide epidemic. In general, to estimate survival, a number of survival time models, such as non-parametric models, like the Kaplan-Meier approach, as well as parametric models like those based on Weibull, Exponential and Gamma distribution plus semiparametric models like the Cox model, are being used to predict mortality and to assess the impact of various predictive factors on patient survival. In many situations however, "off the shelf" methods cannot accurately assess rapid changes in the hazard of mortality experienced by HIV patients starting ART. More recently, Yiannoutsos (Yiannoutsos, 2009) used a Weibull models with a one and two change points to model the high mortality experienced by HIV patients right after the start of therapy. Another possibility, which I explore in this research, is using flexible survival models such as those, for example, developed by Lambert and Royston (Lambert & Royston, 2009), which rely on the use of splines to model more complex behaviors in the data and provide a close representation of the data being analyzed.

The present study uses actual data collected from a sample of people living with HIV, who receive care and treatment in health programs in Uganda. We extend the change-point

approach used by Yiannoutsos (Yiannoutsos, 2009) and compare our spline-based method with the semiparametric Cox model (Lin & Halabi, 2013) with flexible survival models of Lambert and Royston (Lambert & Royston, 2009) to evaluate their effectiveness in modeling the mortality rate of HIV-infected patients.

Chapter Two: Methods

To achieve this objective, this study uses parametric and nonparametric statistical models for fitting and estimating survival time using different covariates. Weibull, exponential, and gamma distributions will be used for the parametric models, while the Cox Model will be used for comparison. The objective of this study is to compare these models to find the best fitting model so that it can then be used to improve modeling of the survival time in HIV-infected individuals starting therapy in this context and explore the pattern of change in mortality rates for a cohort of HIV-infected patients recruited in a care and treatment program in Uganda. Among the class of semiparametric models in particular, we explore the use of flexible spline-based models for survival which expand traditional survival models to reflect changes in survival which may not be fully captured by traditional modeling approaches.

Some Existing Models

The Weibull Model

The Weibull model is a popular model for survival analysis. The model depends on the cumulative distribution function of the Weibull. The hazard function for *T* survival time for the i_{th} patient is given by

$$h_0(t_i) = \left(\frac{\rho}{\lambda}\right)\left(\frac{t_i}{\lambda}\right)^{\rho-1}$$

If covariates are included in the above model then we have

$$h(t_i; z) = \left(\frac{\rho}{\lambda}\right)\left(\frac{t_i}{\lambda}\right)^{\rho-1} e^{\boldsymbol{\beta}'\boldsymbol{z}}$$

Where $\boldsymbol{\beta}$ and $\boldsymbol{z}$ represent vectors of coefficients and covariates respectively. In other words, the effect of $\boldsymbol{z}$ is *multiplicative* on the hazard. The Weibull model for the cumulative hazard up to time t_i is given by the expression

$$H(t_i; z) = \int_0^{t_i} h(u; z)du = \int_0^{t_i} h_o(u)e^{\beta' z}du$$

and we also have

$$H(t_i; z) = H_o(t_i)e^{\boldsymbol{\beta}'\boldsymbol{z}} = \left(\frac{t}{\lambda}\right)^{\rho} e^{\boldsymbol{\beta}' \boldsymbol{z}}$$

The above model is a proportional hazard model because the hazard at time t_i and covariate $\boldsymbol{z}$ is represented as $h(t_i; z) = h_0(t_i)\Psi$ where $\Psi = e^{\boldsymbol{\beta}'\boldsymbol{z}}$ is a constant of proportionality independent of time. The log hazard distribution function is given by

$$\log H(t_i; z) = \rho \log t_i - \rho \log \lambda + \boldsymbol{\beta}' \boldsymbol{z}$$

If $\rho = 1$ then the above model is equivalent to the exponential survival model. The Weibull model is appropriate for many real-life applications and is more flexible in relation to exponential model because it does not assume a constant hazard function of death (see for example Noura & Read, 1990).

The Generalized Gamma Model

The generalized gamma is based on a two-parameter distribution with parameters α and σ. If $\alpha = 1$ then the Gamma model becomes the Weibull model. The survival and hazard distribution of the Gamma model is complex therefore it is not explicitly written here. For further details, see Stacy (1962).

The Cox Model

The Cox model is a famous mathematical model used in medical research and other contexts for survival analysis. It represents the hazard for time *t* as the product of baseline hazard and a function of the covariates.

$$h(t) = h_o(t)e^{\boldsymbol{\beta}'\boldsymbol{z}}$$

Where $\boldsymbol{z}$ are predictor variables. For further details see Kleinbaum & Klein (2005).

An attractive assumption of the above model is that it is a proportional hazards model, since the "baseline" hazard function $h_o(t)$ does not contain any covariate information, while the linear term $\boldsymbol{\beta}'\boldsymbol{z}$ in the exponential expression above is free of time. The Cox model can be extended if variables $\boldsymbol{z}$ are time-dependent, although this technically violates the assumption of proportional hazards. The interpretation of the baseline hazard is the hazard that corresponds to the situation where all $\boldsymbol{z}$ are set to zero. This may not have a physical interpretation. For example, no one has age equal to zero in a problem involving age as a covariate. For further details see Lambert & Royston (2009).

Proposed Models

The Parametric Flexible Survival Models

The parametric models listed above make some strong structural assumptions for the nature of the hazard and, implicitly, the survival distribution. The Cox model does not make such assumptions for the baseline hazard but neither does it model it explicitly. We can also consider models which combine a parametric distribution with a non-parametric estimate of the underlying hazard. Such parametric flexible survival models are implemented for

example in the R package as **flexsurv**. In general, the software fits the probability density of death at time *t* as;

$$f(t|\mu(z),\alpha(z)), \text{where } t \geq 0$$

The cumulative distribution function $F(t)$, the survivor function $S(t) = 1 - F(t)$, the cumulative hazard function $H(t) = -\log S(t)$ and hazard $h(t) = f(t)/S(t)$ are also defined (suppressing the conditioning on the covariate vector z for clarity). In addition, $\mu = \alpha_0$ is the parameter of primary interest, which usually determines the mean of the distribution. Other parameters $\alpha = (\alpha_1, \ldots, \alpha_R)$ are called "ancillary" and determine the shape, variance or higher moments (Jackson, 2016).

The model may sometimes include covariates such that all parameters may depend on a vector of variable ***z*** through link-transformed linear models $g_0(\mu(z)) = \gamma_0 + \beta' z$ and $g_r(\alpha_r(z)) = \gamma_r + \beta' z$. $g(\cdot)$ will typically be the logarithmic function if the parameter is defined to be positive, or the identity function if the parameter is unrestricted (resulting in simple linear regression models). Suppose that the location parameter, but not the ancillary parameters, depends on covariates. If the hazard function factors as $h(t|\alpha, \mu(z)) = \mu(z) h_0(t|\alpha)$, then this is a proportional hazards (PH) model, so that the hazard ratio between groups (defined by different values of *z*) is constant over time *t*. Alternatively, if $S(t|\mu(z), \alpha) = S_0(\mu(z)t|\alpha)$ then it is an accelerated failure time (AFT) model, so that the effect of covariates is to speed or slow the passage of time. We are not concerned with these models in this thesis.

The Royston and Parmar Spline Model

Royston & Parmar (Royston & Parmar, 2002) presented a spline-based survival model, which is a transformation $g(S(t,z))$ of the survival function and it is modelled as a natural cubic spline function of log time: $g(S(t,z)) = s(x,\gamma)$ where $x = \log(t)$. This model can be fitted in **flexsurv** R package using the function **flexsurvspline**. It is also achievable in STATA as presented in a user-supplied STATA program by (Lambert & Royston, 2009). In the R package **flexsurv** by (Jackson, 2016), a log-log transformation approach was used with $g(S(t,z))$ defined as:

$$g\big(S(t,z)\big) = \log(-\log(S(t,z))) = \log(H(t,z)).$$

The last expression $\log(H(t,z))$ is the log of cumulative hazard function in a proportional-hazard model. The spline function $s(x,\gamma)$ is then parameterized as:

$$s(x,\gamma) = \gamma_0 + \gamma_1 x + \gamma_2\, v_1(x) + \cdots + \gamma_{m+1}\, v_m(x)$$

where $v_j(x)$ is the *i*th basis function defined as:

$$v_j(x) = \big(x - k_j\big)_+^3 - \lambda_j (x - k_{min})_+^3 - \big(1 - \lambda_j\big)(x - k_{max})_+^3$$

where $\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}$, $(x-a)_+ = \max(0, x-a)$ and k_j are the knots in the spline function. If $m = 0$ then there are only two parameters γ_0, γ_1, and this is a Weibull model if $g(\cdot)$ is the log cumulative hazard.

Chapter Three: Results

The model comparisons without covariate fitting summaries are presented in Table 1. The goodness of fit measures used are -2log-likelihood and the Akaike Information Criterion (AIC). The best model is the model with the lowest AIC or -2log-likelihood. From Table 1, the best models are proportional odds models with one internal spline knot (scale = "odds") and proportional hazards models with one internal spline knot (scale = "hazard"). The models are better than all other models including all parametric models and the semi-parametric Cox models.

Table 1: Comparison of conventional models with Flexible survival models (without covariates)

Models	**-2 log likelihood**	**Parameters**	**AIC**
Cox	1561.5	0	1561.5
Weibull	1470.8	2	1474.8
Generalized gamma	1445.8	3	1451.7
Log-logistic-Spline	1469.2	2	1473.2
Log-normal-Spline	1459.8	2	1463.8
k=1, Proportional Hazard-Spline	**1440.4**	**3**	**1446.4**
k=1, Proportional Odds-Spline	**1440.4**	**3**	**1446.5**
k=1, Inverse Normal-Spline	1442.4	3	1448.5

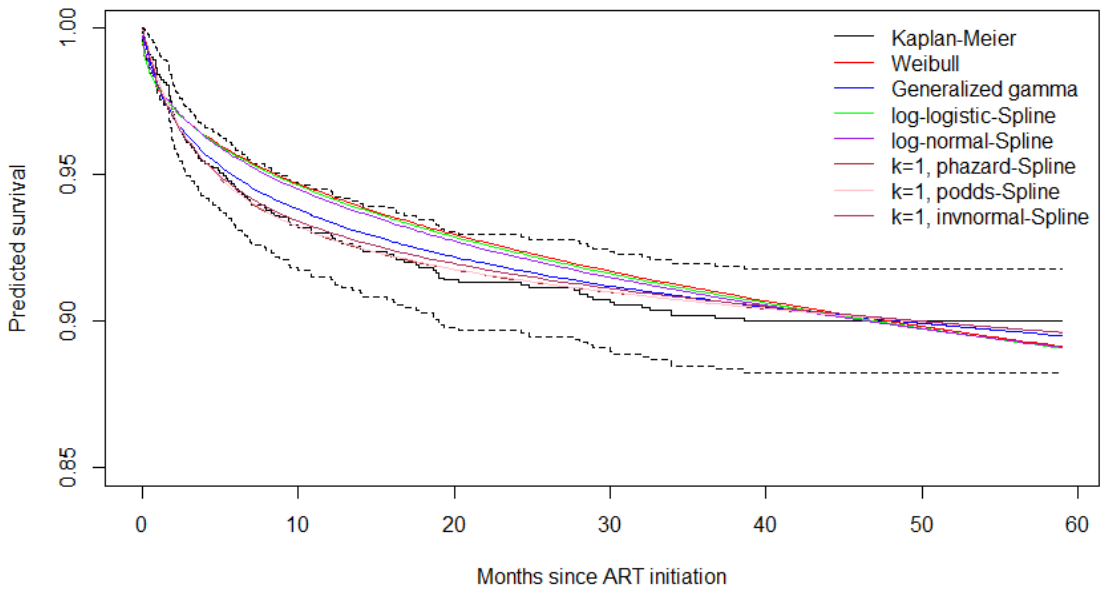

Figure 1: Kaplan Meier (step function) patient survival (smooth curve) versus conventional models with Flexible survival models (without covariates).

Figure 1. shows that the closest survival estimates to the true observed survival estimates (Kaplan Meier) are the proportional hazards models with one internal spline knot (scale = "hazard") and proportional odds models with one internal spline knot (scale = "odds"). The plot also shows that Weibull, Generalized Gamma, log-logistic, lognormal and over (under) estimates the survival (mortality) of the patients immediately following ART initiation.

Table 2 presents the model comparison with covariate fitting summary. The goodness of fit measures used are the -2log-likelihood and Akaike Information Criterion as above. The best model is the model with the lowest AIC or -2log-likelihood. From Table 2, the best model is proportional odds models with one internal spline knot (scale = "odds"). The model is better than all other models especially the off shelf parametric and semi-parametric Cox models.

Table 2: Comparison of off shelf models with Flexible survival models (with covariates)

Models	**-2 log likelihood**	**Parameters**	**AIC**
Cox	1502.9	4	1510.9
Weibull	1414.8	6	1426.9
Generalized gamma	1404.2	7	1418.3
Log-logistic-Spline	1412.4	6	1424.4
Log-normal-Spline	1405.8	6	1417.9
k=1, Proportional Hazard-Spline	1385.2	7	1399.1
k=1, Proportional Odds-Spline	**1384.8**	**7**	**1398.9**
k=1, Inverse Normal-Spline	1390.4	7	1404.4

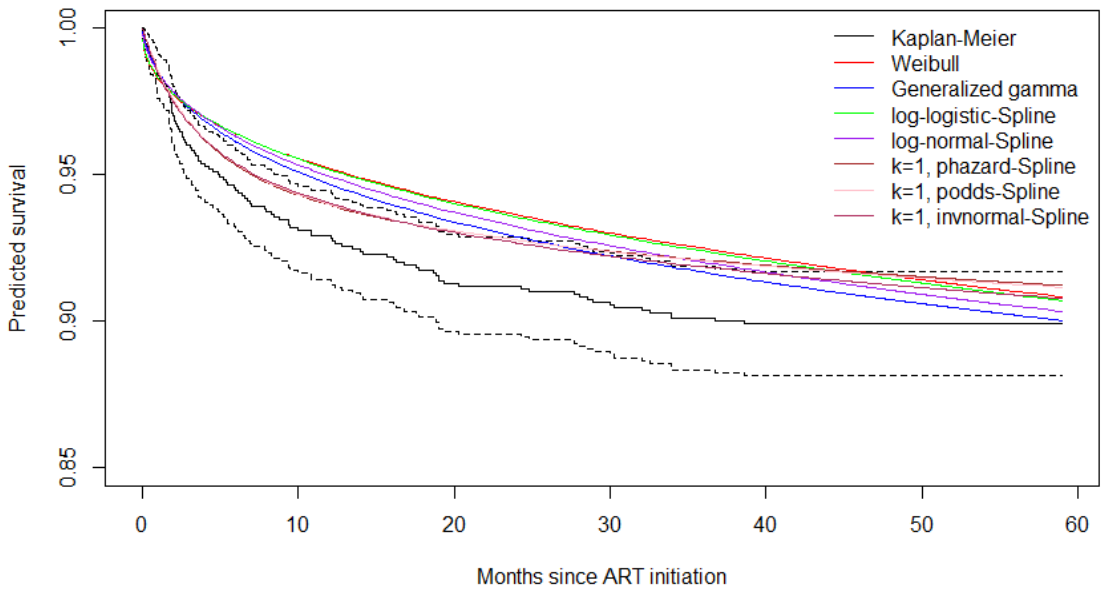

Figure 2: Kaplan Meier (step function) patient survival (smooth curve) versus conventional models with Flexible survival models (with covariates).

Figure 2 shows that the closest survival estimates to the true observed survival estimates (Kaplan Meier) are proportional hazards models with one internal spline knot (scale = "hazard") and proportional odds models with one internal spline knot (scale = "odds"). Furthermore, the plot also shows that the Weibull, Generalized Gamma, log-logistic, log-normal, and the proportional hazards models with one internal spline knot (scale = "hazard"), proportional odds models with one internal spline knot (scale = "odds") and inverse normal models with one internal spline knot (scale = "normal") over (under) estimates the survival (mortality) of the patients immediately following ART initiation.

Table 3: Summary of the best model (proportional odds models with one internal spline knot)

Parameter	Estimate	L95%	U95%	SE	Odds -ratio	L95%	U95%
γ_0	-6.132	-8.444	-3.820	1.180			
γ_1	1.190	0.836	1.544	0.181			
γ_2	0.031	0.018	0.044	0.007			
Age	0.005	-0.019	0.029	0.012	1.005	0.981	1.030
basecd4	-0.006	-0.009	-0.004	0.001	0.994	0.991	0.996
baselogvl	0.547	0.172	0.922	0.191	1.728	1.188	2.514
male	0.192	-0.245	0.628	0.223	1.212	0.783	1.875

Table 3 shows that the effect of age and gender are not significant in the model given that the odds ratio confidence interval includes 1. However, the baseline CD4 counts and baseline log of viral loads are significant. The effect of baseline CD4 shows that increase in baseline CD4 reduces the hazard rate by 0.994 times per higher cell/μl of CD4-positive lymphocyte cells while a one-unit increase in baseline log of viral load increases the hazard rates by 1.728 times.

Chapter Four: Conclusions

The objective of this thesis research was to determine how mortality of HIV-infected patients changes after the initiation of ART therapy. Comparison of various modeling approaches shows that the flexible survival models are superior than a number of conventional parametric and semi-parametric models both in terms of model fitting, comparing visually their fit against that produced by the Kaplan-Meier approach, as well as using the AIC or log-likelihood criteria. We conclude that mortality of HIV-infected patients high after the initiation of ART therapy and decreases rapidly.

References

Jackson, C. H. (2016, May 12). flexsurv: A Platform for Parametric Survival Modeling in R. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5868723/

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis: a self-learning text*. New York: Springer.

Lambert, P. C., & Royston, P. (2009). Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal: Promoting Communications on Statistics and Stata, 9*(2), 265–290. doi: 10.1177/1536867x0900900206

Lin, C.-Y., & Halabi, S. (2013, November 20). On model specification and selection of the Cox proportional hazards model. Retrieved from https://ncbi.nlm.nih.gov/pmc/articles/pmc3795916

Noura, A. A., & Read, K. L. Q. (1990). Proportional Hazards Changepoint Models in Survival Analysis. *Applied Statistics*, *39*(2), 241. doi: 10.2307/2347763

May, M. T., Vehreschild, J.-J., Trickey, A., Obel, N., Reiss, P., Bonnet, F., … Sterne, J. A. C. (2016). Mortality According to CD4 Count at Start of Combination Antiretroviral Therapy Among HIV-infected Patients Followed for up to 15 Years After Start of Treatment: Collaborative Cohort Study. *Clinical Infectious Diseases*, *62*(12), 1571–1577. https://doi.org/10.1093/cid/ciw183.

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine, 21*(15), 2175–2197. doi: 10.1002/sim.1203

Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics, 33*(3), 1187–1192. doi: 10.1214/aoms/1177704481

Stringer, J. S. A., Zulu, I., Levy, J., Stringer, E. M., Mwango, A., Chi, B. H., … Sinkala, M. (2006). Rapid Scale-up of Antiretroviral Therapy at Primary Care Sites in Zambia. *Jama*, *296*(7), 782. doi: 10.1001/jama.296.7.782

Yiannoutsos, C. T. (2009). Modeling AIDS survival after initiation of antiretroviral treatment by Weibull models with changepoints. *Journal of the International AIDS Society*, *12*(1), 9. doi: 10.1186/1758-2652-12-9

Curriculum Vitae

Shaher Bahakeem

Education

- Master of Science in Biostatistics, earned at Indiana University-Purdue University Indianapolis, (July 2020).
- Bachelor of Science in Statistics, earned at King Abdulaziz University, (2015).

Training Experience

- Training Experience: statistics of the number of aircraft during pilgrimage season, at Saudi Civil Aviation (2011-2015)

Professional Experience

- Public Relations Officer – Michelin Co, (2015-2016)

Conferences Attended

- Statistical Analysis Forum Using the SPSS Program (10/28/2014)